Feature Selection Inspired Classifier Ensemble Reduction

Diao, Ren; Chao, Fei; Peng, Taoxin; Snooke, Neal; Shen, Qiang

Published in:
IEEE Transactions on Cybernetics

DOI:
10.1109/TCYB.2013.2281820

Publication date:
2014

Citation for published version (APA):
Diao, R., Chao, F., Peng, T., Snooke, N., & Shen, Q. (2014). Feature Selection Inspired Classifier Ensemble Reduction. *IEEE Transactions on Cybernetics*, *44*(8), 1259-1268. https://doi.org/10.1109/TCYB.2013.2281820

Feature Selection Inspired Classifier Ensemble Reduction

Ren Diao, Fei Chao, Taoxin Peng, Neal Snooke and Qiang Shen

***Abstract*—Classifier ensembles constitute one of the main research directions in machine learning and data mining. The use of multiple classifiers generally allows better predictive performance than that achievable with a single model. Several approaches exist in the literature that provide means to construct and aggregate such ensembles. However, these ensemble systems contain redundant members that, if removed, may further increase group diversity and produce better results. Smaller ensembles also relax the memory and storage requirements, reducing system's run-time overhead while improving overall efficiency. This paper extends the ideas developed for feature selection problems in order to support classifier ensemble reduction, by transforming ensemble predictions into training samples, and treating classifiers as features. Also, the global heuristic harmony search is used to select a reduced subset of such artificial features, while attempting to maximise the feature subset evaluation. The resulting technique is systematically evaluated using high dimensional and large sized benchmark data sets, showing a superior classification performance against both original, unreduced ensembles and randomly formed subsets.**

***Index Terms*—Classifier ensemble reduction, feature selection, harmony search.**

I. Introduction

The main purpose of a classifier ensemble [40], [49] is to improve the performance of single classifier systems. Different classifiers usually make different predictions on certain samples, caused by their diverse internal models. Combining such classifiers has become the natural way of trying to increase the classification accuracy, by exploiting their uncorrelated errors. Also, each ensemble member can potentially be trained using a subset of training samples, which may reduce the computational complexity issue that arises when a single classification algorithm is applied to very large data sets. Additionally, an ensemble can operate in a distributed environment, where data sets are physically separated and are cost ineffective or technically difficult to be combined into one database, in order to train a single classifier. A typical approach to building classifier ensembles involves building a group of classifiers with diverse training backgrounds [5], [19], before combining their decisions together to produce the final prediction. Instead of adopting a simple majority voting-based aggregation [29], methods have also been developed that employ meta-level learners in order to combine the outputs of the base classifiers. Such methods are referred to as "ensemble stacking" in the literature [12].

The target of classifier ensemble reduction (CER) [50] (or classifier ensemble pruning) is to reduce the amount of redundancy in a pre-constructed classifier ensemble, in order to form a much reduced subset of classifiers that can still deliver the same classification results. It is an intermediate step between ensemble construction and decision aggregation. Efficiency is one of the obvious gains from CER. Having a reduced number of classifiers can eliminate a portion of run-time overheads, making the ensemble processing quicker; having fewer models also means relaxed memory and storage requirements. Removing redundant ensemble members may also lead to improved diversity within the group, and further increase the prediction accuracy of the ensemble. Existing approaches in the literature include techniques that employ clustering [16] to discover groups of models that share similar predictions, and subsequently prune each cluster separately. Others use reinforcement learning [41] and multi-label learning [38] to achieve redundancy removal. A number of similar approaches [29], [51] focus on selecting a potentially optimal subset of classifiers, in order to maximise a certain pre-defined diversity measure.

The main aim of feature selection (FS) is to discover a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original data [8], [23]. Practical problems arise when analysing data that have a very large number of features [44], [57], so-called "curse-of-dimensionality" [2], and when it is difficult to identify and extract patterns or rules due to the high inter-dependency amongst individual features, or the behaviour of combined features. Given a data set with n features, the task of FS can be seen as a search for an "optimal" feature subset through the competing 2^n candidate subsets. Optimality is subjective depending on the problem at hand, and a subset that is selected as optimal using one particular evaluator may not be equivalent to that of a subset selected by another. Various techniques have been developed in the literature to judge the quality of the discovered feature subsets, such as methods based on rough sets [35], [48] and fuzzy-rough sets [4], [23], [24], probabilistic consistency [7], and correlation analysis [17]. An unsupervised FS method [34] has also been proposed which operates on un-labelled data. The above mentioned techniques are often referred to as filter-based approaches that are independent of any learning algorithm subsequently employed. In contrast, wrapper-based [20], [27] and hybrid algorithms [60] are often used in conjunction with a learning or data mining algorithm, which is employed in place of an

Ren Diao, Neal Snooke, and Qiang Shen are with the Institute of Mathematics, Physics and Computer Science, Aberystwyth University, UK (email: {rrd09, nns, qqs}@aber.ac.uk). Fei Chao is with the Cognitive Science Department, Fujian Key Laboratory of Brain-like Intelligent System, Xiamen University, Xiamen, Fujian, China (email: fchao@xmu.edu.cn). Taoxin Peng is with the School of Computing, Edinburgh Napier University, 10 Colinton Road, Edinburgh, UK (email: t.peng@napier.ac.uk).

evaluation metric as used in the filter-based approach.

To locate the "optimal" feature subset, an exhaustive method might be used, however it is often impractical for most data sets. Alternatively, hill-climbing based approaches are exploited where features are added or removed one at a time until there is no further improvement to the current candidate solution. Although generally fast to converge, these methods may lead to the discovery of sub-optimal subsets [33], both in terms of the evaluation score and the subset size [11]. To avoid such short-comings, other algorithms utilise random search or nature inspired heuristic strategies such as genetic algorithms [30], [56], simulated annealing [9], and particle swarm optimisation [53] with varying degrees of success. Harmony search (HS) [15], [31] in particular, is a recently developed meta-heuristic algorithm that mimics the improvisation process of music players. It imposes only limited mathematical requirements and is insensitive to initial value settings. Due to its simplistic structure and powerful performance, HS has been very successful in a wide variety of engineering [13], [47] and machine learning tasks [36], [39], [42], and demonstrated several advantages over traditional techniques. HS has been successfully applied to solving FS problems [11], dynamic parameter tuning and iterative solution refinement techniques have also been proposed to further improve the search outcome.

In this paper, a new framework for CER is proposed which builds upon the ideas from existing FS techniques. Inspired by the analogies in between CER and FS, this approach attempts to discover a subset of classifiers by eliminating redundant group members, while maintaining (or increasing) the amount of diversity within the original ensemble. As a result, the CER problem is being tackled from a different angle: each ensemble member is now transformed into an artificial feature in a newly constructed data set, and the "feature" values are generated by collecting the classifiers' predictions. FS algorithms can then be used to remove redundant features (now representing classifiers) in the present context, in order to select a minimal classifier subset while maintaining original ensemble diversity, and preserving ensemble prediction accuracy. The current CER framework extends the original idea [10] that works exclusively with the fuzzy-rough subset evaluator [24], thus allowing many different FS evaluators and subset search methods to be used. It is also made scalable for reducing very large classifier ensembles.

The fusion of CER and FS techniques is of particular significance for problems that place high demands on both accuracy and speed, including intelligent robotics and systems control [32]. For instance, simultaneous mapping and localisation has been recognised to be a very important task for building robots [37]. To perform such tasks, apart from the direct use of raw data or simple features as geometric representations, different approaches that capture more context information have been utilised recently [28]. It has been recognised that ensemble-based methods may better utilise these additional cognitive and reasoning mappings to boost the performance. In effect, CER may be adopted to prune down the redundant, unessential models, so that the complexity of the resultant system is restricted to a manageable level. Also, FS has already been successfully applied to challenging real-world problems like Martian terrain image classification [46], and to reducing the computational costs in vision-based robot positioning [54] and activity recognition [52]. It is therefore, of natural appeal to be able to integrate classifier ensemble and CER to further enhance their potential.

The rest of this paper is laid out as follows. Section II explains the basic structure of HS, and the HS based FS algorithm that serves as the fundamental platform upon which the CER system is developed. Section III introduces the key concepts of the proposed CER framework, illustrates how it can be modelled as an FS problem, and details the approach developed to tackle the problem. Section IV presents the experimentation results along with discussions. Section V concludes the paper and proposes further work in the area.

II. Feature Selection with Harmony Search

HS [15] acts as a meta heuristic algorithm which attempts to find a solution vector that optimises a given (possibly multivariate) cost function. In such a search process, each decision variable (musician) generates a value (note) for finding a global optimum (best harmony). HS has a novel stochastic derivative (for discrete variables) based on musician's experience, rather than gradient (for continuous variables) in differential calculus. This section describes the HS based FS technique (HSFS), and explains how an FS problem can be converted into an optimisation problem, further solved by HS.

A. Key Concepts Mapping

The key concepts of HS are musicians, notes, harmonies and harmony memory. For conventional optimisation problems, the musicians are the decision variables of the cost function being optimised, their values are referred to as playable notes. Each harmony contains notes played by all musicians, or a solution vector containing the values for each decision attribute. The harmony memory holds a selection of played harmonies, which can be more concretely represented by a two dimensional matrix. The number of rows (harmonies) are predefined and bounded by the size of harmony memory (HMS). Each column is dedicated to one musician, it stores the good notes previously played by the musician, and provides the pool of playable notes (referred to hereafter as the note domain) for future improvisations.

When applied to FS, a musician is best described as an independent expert or a "feature selector", and the available features translate to notes. Each musician may vote for one feature to be included in the emerging harmony (feature subset), which is the combined vote from all feature selectors, indicating which features are being nominated. The entire pool of the original features forms the range of notes shared by all musicians. This is different from conventional applications where variables have distinct value ranges. Multiple selectors are allowed to choose the same feature, or they may opt to choose none at all. The fitness function becomes a feature subset evaluator which analyses and merits each of the new subsets found during the search process.

HSFS uses 4 parameters, HMS, the maximum number of iterations K, the number of feature selectors N, and the harmony memory considering rate (HMCR) which encourages the feature selector to randomly choose from all available features (instead of within its own note domain). To lessen the drawbacks lying with the use of fixed parameter values, a dynamic parameter adjustment scheme (for HMS and HMCR) and an iterative refinement procedure (for N) have been proposed to adjust parameter values and improve solution quality. Parameters are dynamically and gradually changed at run time, with different settings being used for the purposes of initial solution space exploration, intermediate solution refinement, and fine tuning towards termination.

TABLE I
HARMONY ENCODED FEATURE SUBSETS

	M_1	M_2	M_3	M_4	M_5	M_6	Represented Subset B
H^1	a_2	a_1	a_3	a_4	a_7	a_{10}	$\{a_1, a_2, a_3, a_4, a_7, a_{10}\}$
H^2	a_2	a_2	a_2	a_3	a_{13}	$-$	$\{a_2, a_3, a_{13}\}$
H^3	a_2	$-$	a_2	$a_3 \rightarrow a_6$	a_{13}	a_4	$\{a_2, a_4, a_6, a_{13}\}$

Table I depicts the following three example harmonies. H^1 denotes a subset of 6 distinctive features: $B_{H^1} = \{a_1, a_2, a_3, a_4, a_7, a_{10}\}$. H^2 shows a duplication of choices from the first three musicians, and a discarded note (represented by $-$) from p^6, representing a reduced subset $B_{H^2} = \{a_2, a_3, a_{13}\}$. H^3 signifies the feature subset $B_{H^3} = \{a_2, a_6, a_4, a_{13}\}$, where $a_3 \rightarrow a_6$ indicates that p^4 originally nominated a_3, but it is forced to change its choice to a_6 due to HMCR activation.

B. Iteration Steps of HSFS

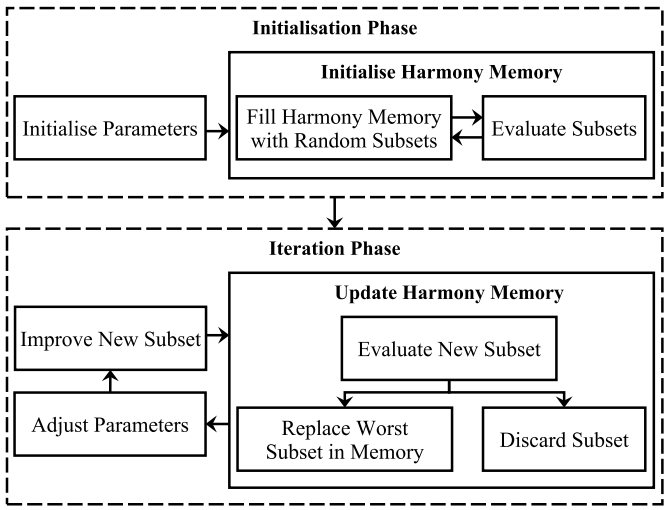


Fig. 1. Parameter Controlled Harmony Search Applied to Feature Selection

The iteration steps of HSFS are demonstrated here using the fuzzy-rough dependency function [24] as the subset evaluator (though other quality metrics may be used as alternatives), accompanied by the flow diagram shown in Fig. 1.

1) **Initialise Problem Domain** The value ranges of the 4 parameters are defined according to the problem domain. The subset storage containing HMS randomly generated subsets is then initialised. This provides each feature selector a working domain of HMS number of features, which may include identical choices, and nulls. The current worst harmony in the memory is hypothetically, $\{a_1, a_2, a_2, a_3, a_6, -\}$ with an evaluation score of 0.5.
2) **Improvise New Subset** A new feature is chosen randomly by each feature selector out of their working feature domain, and together forms a new feature subset: $\{a_1, a_4, a_3, a_3, a_7, -\}$. The 5th selector did not originally have a_7 in its note domain, but the HMCR activation causes it to pick it. This newly emerged subset has an evaluation score of 0.6.
3) **Update Subset Storage** This newly obtained subset achieves a better fuzzy-rough dependency score than that of the worst subset in the subset storage, therefore, the new subset is included in the subset storage and the existing worst subset is removed. The feature a_7 is also introduced to the memory for future combinations. The comparison of subsets takes into consideration both the dependency score and the subset size. This improvisation and update process repeats up to K number of iterations in order to discover the minimal fuzzy-rough reduct (a subset of full fuzzy-rough dependency score) at termination.

HSFS has the strength where that a group of features are being evaluated as a whole. A newly improvised subset does not necessarily get included in the subset storage, just because one of the features has a locally strong fuzzy-rough dependency score. This is the key distinction to any of the hill-climbing based approaches.

III. THE CLASSIFIER ENSEMBLE REDUCTION FRAMEWORK

For most practical scenarios, the classifier ensemble is generated and trained using a set of given training data. For new samples, each ensemble member individually predicts a class label, which are aggregated to provide the ensemble decision. It is inevitable that such ensembles contain redundant classifiers that share very similar if not identical models. This may be caused by the shortage of training data, or the performance limitations of the model diversifying process. Such ensemble members, while occupying valuable system resources, are likely to draw the same class prediction for new samples, therefore provide very limited new information to the group.

The ensemble reduction process, if it occurred in between ensemble generation and aggregation, may reduce the amount of redundancy in the system. The benefit of having a group of classifiers is to maintain and improve the ensemble diversity. The fundamental concept and goals of CER is therefore the same as FS. Having introduced the HSFS technique, the following section aims to explain how a CER problem can be converted into an FS scenario, and details the framework proposed to efficiently perform the reduction. The overall approach developed in this work is illustrated in Fig. 2 containing four key steps.

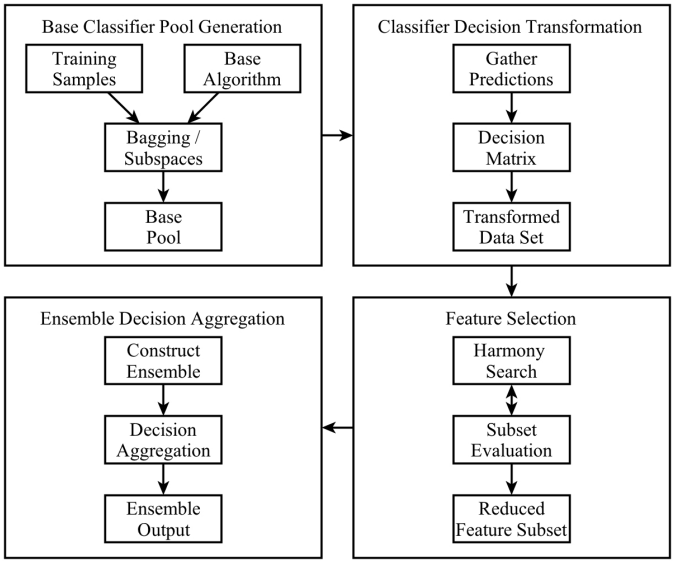


Fig. 2. CER Flow Chart

A. Base Classifier Pool Generation

Forming a diverse base classifier pool (BCP) is the first step in producing a good classifier ensemble. Any preferred methods can be used to build the base classifiers, such as Bagging [5] or Random Subspaces [19]. BCP can either be created using a single classification algorithm, or through a mixture of classifiers. Bagging randomly selects different subsets of training samples in order to build diverse classifiers. Differences in the training data present extra or missing information for different classifiers, resulting in different models. The Random Subspaces method randomly generates different subsets of domain attributes and builds various classifiers on top of each of such subsets. The differences between the subsets creates different view points of the same problem [6], typically resulting in different borders for classification. For a single base classification algorithm, these two methods both provide good diversities. In addition, a mixed classifier scheme is implemented in the presented work. By selecting classifiers from different schools of classification algorithms, the diversity is naturally achieved through the various foundations of the algorithms themselves.

B. Classifier Decision Transformation

TABLE II
DECISION MATRIX

	C_1	C_2	$\cdots$	C_i	$\cdots$	C_{N_C}
I_1	D_{11}	D_{21}	$\cdots$	D_{i1}	$\cdots$	D_{N_C1}
I_2	D_{12}	D_{22}	$\cdots$	D_{i2}	$\cdots$	D_{N_C2}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
I_j	D_{1j}	D_{2j}	$\cdots$	D_{ij}	$\cdots$	D_{N_Cj}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
I_{N_I}	D_{1N_I}	D_{2N_I}	$\cdots$	D_{iN_I}	$\cdots$	$D_{N_CN_I}$

Once the base classifiers are built, their decisions on the training instances are also gathered. For base classifiers $C_i, i = 1, 2, \ldots, N_C$, and training instances $I_j, j = 1, 2, \ldots, N_I$, where N_C is the total number of base classifiers, and N_I is the total number of training instances, a decision matrix as shown in Table II can be constructed. The value D_{ij} represents the ith classifier's decision on the jth instance. For supervised FS, a class label is required for each training sample, the same class attribute is taken from the original data set, and assigned to each of the instances. Note that both the total number of instances and the relations between instances and their class labels remain unchanged. Although all attributes and values are completely replaced by transformed classifier predictions, the original class labels remain the same. A new data set is therefore constructed, each column represents an artificially generated feature, each row corresponds to a training instance, the cell then stores the transformed feature value.

C. Feature Selection on the Transformed Data Set

HSFS is then performed on the artificial data set, evaluating the emerging feature subset using the predefined subset evaluator (such as the fuzzy-rough dependency measure [24]). HSFS optimises the quality of discovered subsets, while trying to reduce subset sizes. When HS terminates, its best harmony is translated into a feature subset and returned as the FS result. The features then indicate their corresponding classifiers that should be included in the learnt classifier ensemble. For example, if the best harmony found by HS is $\{-, C_9, C_3, C_{23}, C_3, C_5, C_{17}, -\}$, the translated artificial feature subset is then $\{C_3, C_5, C_9, C_{17}, C_{23}\}$. Thus, the 3rd, 5th, 9th, 17th and 23rd classifiers will be chosen from the BCP to construct the classifier ensemble.

D. Ensemble Decision Aggregation

Once the classifier ensemble is constructed, new objects are classified by the ensemble members, and their results are aggregated to form the final ensemble decision output. The *Average of Probability* [21] method is used in this paper. Given ensemble members $E_i, i = 1, 2, \ldots, N_E$, and decision classes $D_j, j = 1, 2, \ldots, N_D$, where$N_E$ is the ensemble size and N_D is the number of decision classes, classifier decisions can be viewed as a matrix of probability distributions $\{P_{ij}\}$. Here, P_{ij} indicates prediction from classifier C_i for decision class D_j. The final aggregated decision is the winning classifier that has the highest averaged prediction across all classifiers, as shown in Eq. 1.

$$\{\sum_{i=1}^{N_E} P_{i1}/N_E, \sum_{i=1}^{N_E} P_{i2}/N_E, \ldots, \sum_{i=1}^{N_E} P_{iN_D}/N_E\} \quad (1)$$

Note that this is effective because redundant classifiers are now removed. As such, the usual alternative aggregation method: *Majority Vote* is no longer favourable since the "majority" has now been significantly reduced.

E. Complexity Analysis

Various factors affect the overall complexity of the proposed CER framework, namely the performance of the base classification algorithm and the subset evaluator. Since the proposed

CER framework is generic and not limited to a specific collection of methods, in the following analysis, $O_{C(Train)}$, $O_{C(Test)}$, and O_{Eval} are used to represent the complexity of training and testing the employed base classifier, and that of the subset evaluator, respectively. The amount of time required to construct the base ensemble $(O_{Bagging} + O_{C(Train)}) \times N_C$ can be rather substantial if the size of the ensemble N_C is very large. The process of generating the artificial training data set is straightforward, requiring only $O_{C(Test)} \times N_C \times N_I$, where N_I is the number of instances.

HSFS requires $O_{Eval} \times K$ to perform the subset search, as the total number of evaluations is controlled by the maximum number of iterations K. Note that the subset evaluation itself can be time consuming for high dimensional data (large sized ensembles). As for the complexity of the HS algorithm: the initialisation requires $O(N \times \text{HMS})$ operations to randomly fill the subset storage, and the improvisation process is of the order $O(N \times K)$ because every feature selector needs to produce a new feature at every iteration. Finally, the complexity of predicting the class label for any new sample is $O_{C(Test)} \times N_E$, here N_E is the size of the reduced ensemble.

IV. Experimentation

To demonstrate the capability of the proposed CER framework, a number of experiments have been carried out. The implementation works closely with the WEKA [55] data ming software which provides software realisation of the algorithms employed, and an efficient platform for comparative evaluation. The main ensemble construction method adopted is the Bagging [5] approach, and the base classification algorithm used is C4.5 [55]. The Correlation Based FS [17] (CFS), the Probabilistic Consistency Based FS [7] (PCFS), and the FS technique developed using fuzzy-rough set theory [24] (FRFS) are employed as the feature subset evaluators. The HSFS algorithm then works together with the various evaluators to identify quality feature (classifier) subsets. In order to show the scalability of the framework, the base ensembles are created in 3 different sizes, 50, 100, and 200.

A collection of real-valued UCI [14] benchmark data sets are used in the experiments, a number of which are large in size and high in dimension and hence, present significant challenges to the construction and reduction of ensembles. The parameters used in the experiments and the information of the data sets are summarised in Table III. Stratified 10-fold cross-validation (10-FCV) is employed for data validation. The construction of the base classifier ensemble, and the ensemble reduction process are both performed using the same training fold, so that the reduced subset of classifiers can be compared using the same unseen testing data. The stratification of the data prior to its division into different folds ensures that each class label has equal representation in all folds, thereby helping to alleviate bias/variance problems [3]. The experimental outcomes presented are averaged values of 10 different 10-FCV runs (i.e., 100 outcomes), in order to lessen the impact of random factors within the heuristic algorithms.

TABLE III
HS Parameter Settings and Data Set Information

HMS	**# Musicians**	**HMCR**	**K**
10-20	# features	0.5-1	1000
Data set	**Features**	**Instances**	**Decisions**
arrhythmia	280	452	16
cleveland	14	297	5
ecoli	8	336	8
glass	9	214	6
heart	13	270	2
ionosphere	35	230	2
letter	16	20000	26
libras	91	360	15
magic	10	19020	2
ozone	73	2536	2
secom	591	1567	2
sonar	61	208	2
water	39	390	3
waveform	41	5000	3
wine	14	178	3

A. Reduction Performance for C4.5 Based Ensembles

In this set of experiments, the BCP is built using C4.5 [55] as the base algorithm. Table IV summarises the obtained 3 sets of results for CFS, PCFS, and FRFS respectively, after applying CER, as compared against the results of using: (1) the base algorithm itself, (2) the full base classifier pool, and (3) randomly formed ensembles. Entries annotated in bold indicate that the selected ensemble performance is either statistically equivalent or improved significantly when compared against the original ensemble, using paired *t*-test with two-tailed threshold $p = 0.01$.

Two general observations can be drawn across all three setups: (1) The prediction accuracies of the constructed classifier ensembles are universally superior than that achievable by a single C4.5 classifier. Most of the data sets that revealed the most performance increase are either large in size or high in dimension. This confirms the benefit of employing classifier ensembles. (2) All FS techniques tested demonstrate substantial ensemble size reduction, showing clear evidence of dimensionality reduction.

For the original ensembles of size 50, the CFS evaluator performs very well. In 9 out of 13 tested data sets, CFS achieves comparable or better classification accuracy when compared against the original ensemble. The FRFS evaluator also delivers good accuracies in 5 data sets while having fairly small reduced ensembles. The PCFS only produces equally good solutions for the *cleveland* and *ozone* data sets, however, it has the most noticeable ensemble size reduction power. Better classification performance is achieved by the reduced ensembles for the *cleveland*, *glass*, *letter*, and *water* data sets.

For the medium (100) sized ensembles, both CFS and FRFS produce good results in 7 data sets, however, none of which significantly improves the ensemble classification accuracy. Although PCFS only achieves best performance for the *cleveland* data set, it manages to improve the averaged accuracy by 1.7% across all 10×10 reduced ensembles, with an averaged size of 6.7. Note that for the *ozone* and *sonar* data sets, the reduced ensembles discovered by CFS and FRFS both

TABLE IV
C4.5 CLASSIFICATION ACCURACY RESULTS, BOLD FIGURES INDICATE STATISTICALLY EVUIVALENT OR BETTER PERFORMANCE WHEN COMPARED TO THE UNREDUCED ENSEMBLES

	CFS		PCFS		FRFS		Random		Full		Base C4.5
Data set	Acc.%	Size	Acc.%	Size	Acc.%	Size	Acc.%	Size	Acc.%	Size	Acc.%
Base Ensembles of Size 50											
arrhythmia	**74.59**	**21.6**	71.93	5.3	74.81	26.3	73.71	10	74.47	50	66.39
cleveland	55.54	25.8	**56.57**	**5.7**	**56.60**	**13.6**	54.16	10	54.90	50	50.21
ecoli	**84.55**	**11.6**	83.95	6.8	83.96	23.8	83.94	10	84.24	50	81.88
glass	**74.46**	**15.0**	66.45	4.6	**76.71**	**11.9**	72.94	10	70.24	50	70.15
ionosphere	**91.30**	**10.8**	90.00	3.2	90.43	3.1	90.00	10	90.87	50	87.39
letter	**93.85**	**40.5**	93.10	11	93.54	21.5	92.29	10	93.68	50	87.92
libras	79.44	23.2	74.72	3.5	78.89	15.4	77.78	10	81.67	50	71.39
magic	**87.50**	**29.1**	87.38	38.7	**87.45**	**37.6**	**87.45**	**10**	87.47	50	85.04
ozone	93.88	26.2	**94.12**	**12.3**	**93.96**	**43**	93.40	10	94.00	50	92.94
secom	**93.30**	**35.9**	92.79	6.3	92.92	6.3	**93.11**	**10**	93.24	50	89.28
sonar	75.31	24.5	71.93	3.3	71.05	3.2	72.45	10	75.88	50	70.05
water	**87.69**	**20.9**	83.33	4	84.61	6.1	84.87	10	86.67	50	80.00
waveform	**82.92**	**42.2**	81.50	8.7	**82.47**	**11**	81.00	10	82.98	50	75.50
Base Ensembles of Size 100											
arrhythmia	73.91	28.3	73.04	5.2	**74.37**	**22.3**	73.26	20	74.47	100	66.39
cleveland	54.56	30.4	**58.26**	**6.7**	54.46	11.8	55.56	20	56.56	100	50.21
ecoli	84.85	13.7	85.76	6.4	85.16	24.2	84.84	20	84.25	100	81.88
glass	71.60	16.5	70.58	4.5	**74.31**	**11.7**	72.53	20	74.42	100	70.15
ionosphere	89.13	14.3	90.43	3.1	84.35	3.2	90.87	20	91.74	100	87.39
letter	**93.99**	**58.6**	93.23	11	**93.58**	**27.6**	93.21	20	93.66	100	87.92
libras	**80.83**	**33.0**	74.17	3.5	77.78	15.3	77.22	20	80.28	100	71.39
magic	**87.56**	**38.1**	87.44	38.9	**87.56**	**40.1**	87.32	20	87.56	100	85.04
ozone	**94.24**	**31.8**	93.84	13.5	**94.16**	**74.2**	**94.16**	**20**	94.16	100	92.94
secom	**93.43**	**59.4**	93.04	6.2	92.51	6.1	93.00	20	93.30	100	89.28
sonar	**75.36**	**30.4**	72.88	3.8	**75.36**	**3.5**	72.93	20	75.36	100	70.05
water	**87.18**	**25.7**	85.64	4.7	86.15	6.2	**87.18**	**20**	86.92	100	80.00
waveform	83.20	71	80.88	9	**83.33**	**11**	82.90	20	83.42	100	75.50
Base Ensembles of Size 200											
arrhythmia	**75.47**	**39.9**	72.80	5.7	73.04	21.3	74.37	40	75.25	200	66.39
cleveland	**57.93**	**45**	52.56	5.8	**55.24**	**11.9**	**55.54**	**40**	54.90	200	50.21
ecoli	83.96	24.5	83.94	6.6	84.29	24.3	**84.86**	**40**	84.54	200	81.88
glass	72.53	25.9	72.97	4.8	72.49	11.6	72.08	40	73.94	200	70.15
ionosphere	90.87	20	86.09	3.2	90.87	3.6	89.57	40	91.74	200	87.39
letter	**94.17**	**72.8**	93.33	10.6	**93.71**	**32.07**	93.56	40	93.82	200	87.92
libras	**81.67**	**41**	74.17	4	**81.11**	**15.1**	79.17	40	79.44	200	71.39
magic	**87.64**	**44.8**	**87.62**	**36.5**	87.52	41.8	87.34	40	87.63	200	85.04
ozone	**94.55**	**45**	93.65	28.9	**94.49**	**143**	**94.40**	**40**	94.24	200	92.94
secom	**93.36**	**95.7**	92.92	6.2	**93.38**	**6**	**93.30**	**40**	93.36	200	89.28
sonar	**78.69**	**45.8**	73.36	4.1	74.31	4.5	74.88	40	75.83	200	70.05
water	**87.95**	**38.6**	83.33	4.3	85.87	6.8	86.67	40	87.95	200	80.00
waveform	**83.12**	**107.2**	81.06	9.3	82.40	12	82.76	40	83.48	200	75.50
Average	82.72	36.5	80.89	9.2	82.03	20.9	81.81	23.3	82.63	116.7	77.55
Equal/better	26/39		4/39		17/39		7/39		-		-

show very similar averaged accuracy, which is almost identical to that of the original full ensembles. This may indicate that the key ensembles members are certainly present in the reduced subsets, while FRFS eliminates the most redundancy (average reduced ensemble size is 3.5) for the *sonar* data set.

For the large sized ensembles, CFS shows clear lead in terms of the overall quality of the reduced ensembles, scoring equal classification accuracy for 6 data sets, and delivers an improvement in ensemble accuracy for the *cleveland*, *letter*, *libras*, and *sonar* data sets. This experimentally demonstrates the capability and benefit of employing the proposed CER framework in dealing with large sized ensembles with large, complex data sets. Note that CFS not only picks important features (correlated with the class), but removes redundant ones (inter-correlated with other features). This characteristic may have contributed to the identification of higher quality features (classifiers). For several data sets, the sizes of the ensembles reduced using CFS are considerably larger than those obtained by the other two evaluation measures. This may have led to the observed ensemble performance. FRFS also produces good quality ensembles with much reduced size, showing its strength in redundancy removal. PCFS is not competitive in this set of experiment, this may have been caused by its (perhaps overly) aggressive reduction behaviour, which may have resulted in certain quality ensemble members being ignored.

TABLE V
MIXED CLASSIFIERS USING BAGGING

Data set	Full	FRFS	U-FRFS
cleveland	54.94	52.92 (11.2)	54.27 (9.32)
ecoli	87.67	86.66 (15.98)	85.77 (8.7)
glass	71.12	69.62 (12.2)	69.62 (9.2)
heart	82.07	75.40 (8.76)	77.62 (8.42)
ionosphere	87.73	88.17 (8.36)	88.17 (8.56)
sonar	80.96	73.55 (8.5)	80.76 (8.68)
water	78.15	78.71 (9.26)	78.20 (8.72)
wine	98.31	97.52 (7.74)	97.40 (7.46)

TABLE VI
MIXED CLASSIFIERS USING RANDOM SUBSPACES

Data set	Full	FRFS	U-FRFS
cleveland	56.57	57.85 (11.82)	57.10 (9.08)
ecoli	79.17	84.64 (12.16)	84.40 (7.8)
glass	75.61	71.50 (11.28)	73.08 (8.18)
heart	82.44	80.89 (7.96)	80.44 (8.16)
ionosphere	89.30	87.39 (8.1)	88.00 (7.58)
sonar	82.69	86.06 (7.86)	83.17 (7.88)
water	80.26	80.41 (9.16)	80.92 (8.04)
wine	98.09	97.53 (7.82)	97.75 (7.46)

B. Alternative Ensemble Construction Approaches

For the following set of experiments comparing supervised (FRFS) against un-supervised [34] (U-FRFS) FS approaches, a total of 10 different base algorithms are selected, one or two distinctive classifiers from each representative classifier groups, including fuzzy-based Fuzzy Nearest Neighbour [26] (FNN), Fuzzy-Rough Nearest Neighbour [22], Vaguely Quantified Fuzzy-Rough NN [22], lazy-based IBk [1], tree-based C4.5 [55], REPTree [55], rule-based JRip [55], PART [55], Naive Bayes [25] and Multilayer Perceptron [18]. Bagging and Random Subspaces [19] are then used to create differentiation between classifiers to fill the total BCP of 50. Tables V and VI show the experimental results, using these two methods respectively. Due to the considerable system resource required to construct and maintain the base ensembles, and the complexity invovled in fuzzy-rough set-based feature subset evaluation, this set of experimentations are carried out using ensembles of size 50 with lower dimension benchmark data sets.

For mixed classifiers created using Bagging, the FRFS method find ensembles with much greater size variation. For the *ecoli* data set in particular, the averaged ensemble size is 15.98. The results indicate that many distinctive features (i.e. good diversity classifiers) are present, This particular ensemble also results in the highest accuracy for *ecoli* compared against other approaches, with 87.67% BCP accuracy, and 86.66% ensemble accuracy. A large performance decrease is also noticed for the *sonar* data set. Interestingly, the unsupervised FRFS achieves better overall performance than its supervised counterpart, with smaller selected ensemble sizes.

The Random Subspaces based mixed classifier scheme produces better base pools in 7 out of 9 cases. Both FRFS and U-FRFS find smaller ensembles on average than the case where Bagging is used. Neither method suffers from extreme performance decrease following reduction unlike the results obtained when a single base algorithm is employed. Despite having a BCP that under performs for the *ecoli* data set, both methods manage to achieve an increase of 5% in accuracy. The quality of the mixed classifier group is lower than that of the C4.5 based single algorithm approach for several data sets. This is largely caused by the employment of non-optimised base classifiers. It can be expected that the results achievable after optimisation would be even better.

C. Discussion

Although the execution time of the experimented approaches have not been precisely recorded and presented, it is observed during the study that data sets with large number of instances such as the *ozone*, *secom*, and *waveform* data sets, all require substantial amount of time for the reduction process. This observation agrees with the complexity analysis in Section III-E: the reduction process relies on the efficacy of the evaluators (which may not scale linearly with the number of training instances), and thus, for huge data sets, it may be beneficial to choose the lighter weight evaluators (such as CFS). However, since the reduction process itself can be performed independently and separately from the main ensemble prediction process, CER is generally treated as a preprocessing step (similar to FS) for the ensemble classification, or a post-processing refinement procedure for the generated raw ensembles. The time complexity for such process is less crucial and has less impact.

The experimental evaluation also reveals that different evaluators show distinctive characteristics when producing the reduced ensemble. For example, PCFS consistently delivers very compact ensembles (with less than 10 members for most data sets). CFS excels in terms of ensemble classification accuracy but with much larger sized subsets. FRFS is balanced between ensemble accuracy and dimensionality reduction, with very occasional large solutions (the *ozone* data set). The un-supervised method also produces comparable results to its supervised counterparts.

Note that for a number of experimented data sets, performing CER does not always yield subsets with equal or better performance. This might have been caused by the employed filter-based FS approaches (which do not cross-examine against the original data in terms of classification accuracy). How concepts developed by existing wrapper-based and hybrid FS techniques may be applied to further improve the framework remains active research. The information lost (even the redundant classifiers) through reduction may also be the cause for such decrease in performance. Similar behaviours have also been observed in the FS problem domain. The quality (such as size and variance) of the training data also plays a very important role in CER, the classifiers that were deemed redundant by the subset evaluators may in fact carry important internal models, which are just not sufficiently reflected by the available training samples.

V. CONCLUSION

This paper has presented a new approach to CER. It works by applying FS techniques to minimising redundancy in an artificial data set, generated via transforming a given classifier

ensemble's decision matrix. The aim is to further reduce the size of an ensemble, while maintaining and improving classification accuracy and efficiency. Experimental comparative studies show that several existing FS approaches can entail good solutions by the use of the proposed approach. Reduced ensembles are found with comparable classification accuracies as the original ensembles, and in most cases provide good improvements over the performance achievable by the base algorithms. The characteristics of the results also vary depending on the employed FS evaluator.

Although promising, much can be done to further improve the potential of the presented work. Of particular interest to the authors is the formulation of alternative decision matrix transformation procedures. Many state-of-the-art classifiers are capable of producing a likelihood distribution that a particular instance may belong to the available classes, and the class with highest probability is usually taken as the final prediction. This probability distribution may contain more information, and is potentially more suitable to be used as the artificial feature values (instead of the final prediction). In addition, other statistical information from the classifiers such as variance, may also be good candidates for use as part of the artificially generated features, in order to create a more comprehensive data set for FS. Further experimental evaluation of this work on substantially larger practical problems, such as Martian rock classification [45], [46] and weather forecasting [43], remains as active research. This will help to better understand and validate the characteristics of the employed feature subset evaluators. Investigations are also necessary into the underlying reasons why different FS techniques deliver distinctive characteristics, in either simplifying the complexity of the leant ensembles, or improving overall ensemble prediction accuracy.

Finally, it is worth noting that, instead of the feature subset evaluators adopted in this paper, the proposed approach can be readily generalised to work with other FS techniques (such as feature importance ranking methods), and with alternative heuristic search strategies. The reduction process may be performed in conjunction with diversity enhancing methods such as ensemble selection [40], [49], making the final solution diverse as well as compact. The approach presented in this paper is conceptually similar to that taken for the development of parsimonious fuzzy models [58], [59]. It would be beneficial to have a closer examination in the underlying ideas of such similar work, in an effort to build compact systems of improved generalisation and interpretability.

ACKNOWLEDGMENTS

This work was partly supported by the National Natural Science Foundation of China: No. 61203336, and also by Aberystwyth University and Edinburgh Napier University.

REFERENCES

[1] D. Aha, D. Kibler, and M. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, 1991.

[2] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton University Press, 1957.

[3] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, Sep. 2004.

[4] R. B. Bhatt and M. Gopal, "On fuzzy-rough sets approach to feature selection," *Pattern Recognition Letters*, vol. 26, pp. 965–975, 2005.

[5] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[6] C. M. Christoudias, R. Urtasun, and T. Darrell, "Multi-view learning in the presence of view disagreement," in *Proc. 24th Conference on Uncertainty in Artificial Intelligence*, 2008.

[7] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artificial Intelligence*, vol. 151, no. 1-2, pp. 155–176, Dec. 2003.

[8] ——, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, pp. 131–156, 1997.

[9] J. Debuse and V. Rayward-Smith, "Feature subset selection within a simulated annealing data mining algorithm," *Journal of Intelligent Information Systems*, vol. 9, pp. 57–81, 1997.

[10] R. Diao and Q. Shen, "Fuzzy-rough classifier ensemble selection," in *IEEE International Conference on Fuzzy Systems*, june 2011, pp. 1516 –1522.

[11] ——, "Feature selection with harmony search," *IEEE Trans. Syst., Man, Cybern. B*, vol. 42, no. 6, pp. 1509–1523, 2012.

[12] S. Džeroski and B. Ženko, "Is combining classifiers better than selecting the best one," *Machine Learning*, vol. 54, no. 3, pp. 255–273, Mar. 2004.

[13] M. Fesanghary, M. Mahdavi, M. Minary-Jolandan, and Y. Alizadeh, "Hybridizing harmony search algorithm with sequential quadratic programming for engineering optimization problems," *Computer Methods in Applied Mechanics and Engineering*, vol. 197, no. 33-40, pp. 3080–3091, 2008.

[14] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.

[15] Z. W. Geem, Ed., *Recent Advances In Harmony Search Algorithm*, ser. Studies in Computational Intelligence. Springer, 2010, vol. 270.

[16] G. Giacinto and F. Roli, "An approach to the automatic design of multiple classifier systems," *Pattern Recognition Letters*, vol. 22, pp. 25–33, 2001.

[17] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann, 2000, pp. 359–366.

[18] S. Haykin, *Neural Networks: A Comprehensive Foundation*, ser. International edition. Prentice Hall International, 1999.

[19] T. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[20] C.-N. Hsu, H.-J. Huang, and D. Schuschel, "The annigma-wrapper approach to fast feature selection for neural nets," *IEEE Trans. Syst., Man, Cybern. B*, vol. 32, no. 2, pp. 207–212, 2002.

[21] R. A. Jacobs, "Methods for combining experts' probability assessments," *Neural Computation*, vol. 7, no. 5, pp. 867–888, September 1995.

[22] R. Jensen and C. Cornelis, "Fuzzy-rough nearest neighbour classification," in *Transactions on Rough Sets XIII*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6499, pp. 56–72.

[23] R. Jensen and Q. Shen, *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches*. Wiley-IEEE Press, 2008.

[24] ——, "New approaches to fuzzy-rough feature selection," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 4, pp. 824–838, Aug. 2009.

[25] G. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995, pp. 338–345.

[26] J. Keller, M. Gray, and J. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 4, pp. 580–585, 1985.

[27] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[28] B. Kröse, N. Vlassis, R. Bunschoten, and Y. Motomura, "A probabilistic model for appearance-based robot localization," in *In First European Symposium on Ambience Intelligence (EUSAI*. Springer, 2000, pp. 264–274.

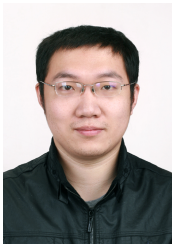
[29] L. Kuncheva, "Switching between selection and fusion in combining classifiers: an experiment," *IEEE Trans. Syst., Man, Cybern. B*, vol. 32, no. 2, pp. 146–156, 2002.

[30] R. Leardi, R. Boggia, and M. Terrile, "Genetic algorithms as a strategy for feature selection," *Journal of Chemometrics*, vol. 6, no. 5, pp. 267–281, 1992.

[31] K. S. Lee and Z. W. Geem, "A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice," *Computer Methods in Applied Mechanics and Engineering*, vol. 194, no. 36-38, pp. 3902–3933, Sep. 2005.

[32] X. Li and L. Parker, "Design and performance improvements for fault detection in tightly-coupled multi-robot team tasks," in *Proceedings of IEEE International Conference on Robotics and Automation*, 2009.
[33] H. Liu and H. Motoda, *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 2007.
[34] N. Mac Parthaláin and R. Jensen, "Measures for unsupervised fuzzy-rough feature selection," *International Journal of Hybrid Intelligent Systems*, vol. 7, no. 4, pp. 249–259, Dec. 2010.
[35] N. Mac Parthaláin, Q. Shen, and R. Jensen, "A distance measure approach to exploring the rough set boundary region for attribute reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 3, pp. 305–317, Mar. 2010.
[36] M. Mahdavi, M. H. Chehreghani, H. Abolhassani, and R. Forsati, "Novel meta-heuristic algorithms for clustering web documents," *Applied Mathematics and Computation*, vol. 201, no. 1-2, pp. 441–451, 2008.
[37] A. Marán-Hernández, R. Méndez-Rodríguez, and F. Montes-González, "Significant feature selection in range scan data for geometrical mobile robot mapping," in *Proceedings of the 5th International Symposium on Robotics and Automation*, 2006.
[38] F. Markatopoulou, G. Tsoumakas, and L. Vlahavas, "Instance-based ensemble pruning via multi-label classification," in *Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on*, vol. 1, 2010, pp. 401–408.
[39] M. H. Mashinchi, M. A. Orgun, M. Mashinchi, and W. Pedrycz, "A tabu-harmony search-based approach to fuzzy linear regression," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 3, pp. 432–448, 2011.
[40] L. Nanni and A. Lumini, "Ensemblator: An ensemble of classifiers for reliable classification of biological data," *Pattern Recognition Letters*, vol. 28, no. 5, pp. 622–630, 2007.
[41] I. Partalas, G. Tsoumakas, and I. Vlahavas, "Pruning an ensemble of classifiers via reinforcement learning," *Neurocomputing*, vol. 72, no. 7–9, pp. 1900–1909, 2009.
[42] C. C. Ramos, A. N. Souza, G. Chiachia, A. X. F. ao, and J. ao P. Papa, "A novel algorithm for feature selection using harmony search and its application for non-technical losses detection," *Computers & Electrical Engineering*, vol. 37, no. 6, pp. 886–894, 2011.
[43] S. Royston, J. Lawry, and K. Horsburgh, "A linguistic decision tree approach to predicting storm surge," *Fuzzy Sets and Systems*, vol. 215, no. 0, pp. 90 – 111, 2013, ¡ce:title¿Theme : Clustering¡/ce:title¿.
[44] M. Shah, M. Marchand, and J. Corbeil, "Feature selection with conjunctions of decision stumps and learning from microarray data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 174–186, 2012.
[45] C. Shang and D. Barnes, "Fuzzy-rough feature selection aided support vector machines for mars image classification," *Computer Vision and Image Understanding*, vol. 117, no. 3, pp. 202 – 213, 2013.
[46] C. Shang, D. Barnes, and Q. Shen, "Facilitating efficient mars terrain image classification with fuzzy-rough feature selection," *International Journal of Hybrid Intelligent Systems*, vol. 8, no. 1, pp. 3–13, Jan. 2011.
[47] R. Srinivasa Rao, S. V. L. Narasimham, M. Ramalinga Raju, and A. Srinivasa Rao, "Optimal network reconfiguration of large-scale distribution system using harmony search algorithm," *IEEE Trans. Power Syst.*, vol. 26, no. 3, pp. 1080–1088, 2011.
[48] R. W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," *Pattern Recognition Letters*, vol. 24, no. 6, pp. 833 – 849, 2003.
[49] M. A. Tahir, J. Kittler, and A. Bouridane, "Multilabel classification using heterogeneous ensemble of multi-label classifiers," *Pattern Recognition Letters*, vol. 33, no. 5, pp. 513 – 523, 2012.
[50] G. Tsoumakas, I. Partalas, and I. Vlahavas, "A taxonomy and short review of ensemble selection," in *Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications*, 2008.
[51] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Diversity in search strategies for ensemble feature selection," *Information Fusion*, vol. 6, no. 1, pp. 83–98, 2005.
[52] D. L. Vail and M. M. Veloso, "Feature selection for activity recognition in multi-robot domains," in *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 2008, pp. 1415–1420.
[53] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognition Letters*, vol. 28, no. 4, pp. 459 – 471, 2007.
[54] G. Wells and C. Torras, "Assessing image features for vision-based robot positioning," *Journal of Intelligent and Robotic Systems*, vol. 30, no. 1, pp. 95–118, 2001.
[55] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., ser. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Jun. 2005.
[56] J. Wróblewski, "Ensembles of classifiers based on approximate reducts," *Fundamenta Informaticae*, vol. 47, no. 3-4, pp. 351–360, Oct. 2001.
[57] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," in *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann, 2001, pp. 601–608.
[58] S.-M. Zhou and J. Q. Gan, "Constructing accurate and parsimonious fuzzy models with distinguishable fuzzy sets based on an entropy measure," *Fuzzy Sets and Systems*, vol. 157, no. 8, pp. 1057 – 1074, 2006.
[59] S.-M. Zhou and J. Gan, "Constructing L2-SVM-based fuzzy classifiers in high-dimensional space with automatic model selection and fuzzy rule ranking," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 3, pp. 398–409, 2007.
[60] Z. Zhu, Y.-S. Ong, and M. Dash, "Wrapper-filter feature selection algorithm using a memetic framework," *IEEE Trans. Syst., Man, Cybern. B*, vol. 37, no. 1, pp. 70–76, 2007.

Ren Diao received the B.A. and M.A. degree in Computer Science from the University of Cambridge, U.K. He is currently a Research Fellow at the Institute of Mathematics, Physics and Computer Science at Aberystwyth University, as a member of the Advanced Reasoning Group. His research interests include fuzzy set theory, nature-inspired heuristics, and machine learning.



Fei Chao received the B.Eng. degree in mechanical engineering from Fuzhou University, China, and the MSc. Degree with distinction in computer science from the University of Wales, U.K., in 2004 and 2005, respectively, and the Ph.D. degree in robotics from Aberystwyth University, Wales, U.K. in 2009. He was a research associate at Aberystwyth University from 2009 to 2010. He is currently an Assistant Professor with the Cognitive Science Department, at Xiamen University, China. He has published around 20 peer-reviewed papers. His research interests include developmental robotics, machine learning, and optimization algorithms. He is a member of IEEE.

Taoxin Peng received his PhD in computer science from the University of Greenwich, London, UK in 2000. He joined the Department of Artificial Intelligence at the University of Edinburgh, as a Research Associate in 1998. Since 1999, he became a Lecturer in the School of Computing at Edinburgh Napier University. His research interests include data quality, data cleaning, data mining and data warehousing, model-based reasoning, temporal reasoning and knowledge representation. His research findings have been published in both peer reviewed international conferences and journals. Dr. Peng is a Fellow of Higher Education Academy (HEA), UK.

Neal Snooke received a University of Wales BSc Honours degree in Microelectronics and Computing in 1990, followed by a Ph.D in Wavelet based image compression in 1994 from Aberystwyth University, UK. He has been research director of a successful spin-off company specializing in automotive electrical FMEA, and is currently a Lecturer at the Institute of Mathematics, Physics and Computer Science at Aberystwyth University specializing in network technologies, ubiquitous computing, and software engineering. His research interests include model-based reasoning, qualitative reasoning, and software analysis with application to automated design analysis tools for electrical, electronic, network-based, and embedded systems. He was also a member of the steering committee for the European Network of Excellence on Qualitative and Model-based reasoning. He has authored over 50 peer-reviewed papers.

Qiang Shen received the Ph.D. degree from Heriot-Watt University, Edinburgh, U.K. and the DSc degree from Aberystwyth University, Wales, U.K. He holds the established chair in computer science and is the Director of the Institute of Mathematics, Physics and Computer Science at Aberystwyth University, U.K. He is a Fellow of the Learned Society of Wales. His research interests include computational intelligence, reasoning under uncertainty, pattern recognition, data mining, and real-world applications of such techniques for intelligent decision support (e.g., crime detection, consumer profiling, systems monitoring, and medical diagnosis). Prof. Shen is a long-serving Associate Editor of two premier IEEE TRANSACTIONS (IEEE TRANSACTIONS ON CYBERNETICS and IEEE TRANSACTIONS ON FUZZY SYSTEMS) and an editorial board member of several other leading international journals. He has authored 2 research monographs and over 300 peer-reviewed papers, including one which received an Outstanding Transactions Paper Award from IEEE.